BMEG3105 Data analytics for personalized genomics and precision medicine

Scribing

Lecture 18: ‘Visualization’

Name: FUNG Cho Ching	SID: 1155203141
Course Lecturer: Yu LI (李煜) Liyu95.com, liyu@cse.cuhk.edu.hk	

What is Single-Cell Analysis?

- **Definition:** Single-cell analysis is a technique that sequences the genetic information (DNA/RNA) from *individual cells*, unlike traditional methods that analyse a bulk population of cells.
- **Why it's Important:** It allows scientists to see the differences between individual cells, which helps to:
    - **Define Heterogeneity:** Understand the diversity within a cell population.
    - **Identify Rare Cell Populations:** Find small, unique groups of cells that might be missed in bulk analysis.
    - **Study Cell Population Dynamics:** Observe how cells change and transition between states.

The Gene Expression Matrix

- This is the fundamental data structure in single-cell RNA sequencing (scRNA-seq).
- **Structure:** It's an $N \times M$ matrix.
    - **N (Rows):** Represents individual cells, each with a unique barcode (UMI).
    - **M (Columns):** Represents genes (around 20,000).
    - **Values (X):** The "counts" of RNA molecules from each gene in each cell.
- **Normalization:** Raw counts are normalized to **CPM (Counts Per Million)** to make cells with different total RNA amounts comparable. The formula is: $\text{CPM}_i = \frac{x_i}{\sum x} \cdot 10^6$.

Challenges in Single-Cell Data Analytics

- Common problems with scRNA-seq data and the initial steps to fix them.
- Challenges:
    - Noise: Technical and biological variability.
    - Doublet: Two or more cells mistakenly sequenced as one.
    - Dropout: A gene is expressed in a cell but not detected (count is zero).
    - Batch Effect: Technical differences between experiments done at different times or by different people.

- Pre-processing Pipeline: The data goes through several cleaning steps: Raw Data -> Quality Control -> Normalization -> Data Correction (e.g., for batch effects) -> Visualization & Feature Selection.

Dimension Reduction - PCA

- **PCA (Principal Component Analysis)** is a common linear dimensionality reduction technique.
- **How it works:** It finds new axes (principal components) that capture the most variance (spread) in the data. The first PC captures the most variance, the second captures the next most, and so on.

Problem of PCA

- The main issue is that PCA, being a linear method, can fail to preserve the original clustering of data, especially when the clusters have complex, non-linear shapes in high dimensions.

Introducing t-SNE

- **t-SNE (t-distributed Stochastic Neighbor Embedding)** is a non-linear dimensionality reduction technique designed specifically for visualization.
- **Goal:** Model the high-dimensional data in a low-dimensional space (2D/3D) so that **similar cells are near each other** and **dissimilar cells are far apart**.
- **The Process (Simplified):**
    1. **Random Initialization:** Points are placed randomly in 2D.
    2. **Iterative Update:** The algorithm moves points around in small steps.
    3. **Attraction & Repulsion:** Points from the same cluster in high-D attract each other in 2D; points from different clusters repel each other.
    4. **Convergence:** The process stops when the positions stabilize.

PCA vs. t-SNE

- This shows a practical example using handwritten digits (each digit is a point in a 784-dimensional space, from 28x28 pixel images).
- **PCA (First 2D):** The clusters of digits (0-9) are often overlapping and not well separated.
- **t-SNE (Right):** The clusters are much more distinct and well-separated, clearly showing its superiority for visualizing cluster structure.

Disadvantages of t-SNE

- **Computational Cost:** It's iterative and can be slow for large datasets.
- **Non-Deterministic:** Different runs can produce different-looking plots.
- **Global Structure Not Preserved:** Distances between clusters in the t-SNE plot may not reflect their true distances.
- **Noisy Patterns:** Can sometimes create patterns that aren't real.
- **UMAP:** Mentioned as a modern alternative that is often faster and can better preserve some global structures.

Single-Cell RNA-seq Analysis Pipeline

  - **Clustering:** Grouping cells based on gene expression similarity.
  - **Trajectory Inference:** Modeling dynamic processes like cell differentiation.
  - **Differential Expression:** Finding genes that are expressed differently between conditions or cell types.

Protein Binding Preference & Experiment

- **Key Idea:** Proteins like Transcription Factors (TFs) and RNA-Binding Proteins (RBPs) don't bind to random sequences; they have a **preference** for specific DNA/RNA sequences.
- **How to Find the Motif:** An experimental method (like a pulldown assay) where a tagged protein is used to isolate all the RNA/DNA fragments it binds to. These fragments are then sequenced and analyzed to find the common binding sequence pattern, or **"motif."**

What is a Motif?

- **Motif:** A conserved, short sequence pattern that represents the preferred binding site for a protein.
- **From Sequences to a Motif:**
  1. **Align** the bound sequences.
  2. Create a **Position Count Matrix (PCM):** Count how often each nucleotide (A,C,G,T) appears at each position in the aligned sequences.
  3. Convert to a **Position Probability Matrix (PPM):** Convert the counts into probabilities (each column sums to 1).

4. Visualize as a **Sequence Logo:** The height of the letters represents how conserved that position is; taller stacks mean a stronger preference for that nucleotide.

Why Do We Care About Health Data?

- **The Data Spectrum:** Health data ranges from molecular (genomics, proteomics) to organ-level (medical imaging) to personal and environmental (diet, lifestyle).
- **For Doctors:** This data is crucial for precise diagnosis (like identifying a snake species from its features) and treatment.
- **The Future: AI + Health Data:** AI can assist doctors by analyzing this vast amount of data to improve disease diagnosis and treatment planning.

AI vs. ML vs. DL

- **Artificial Intelligence (AI):** The broadest field. Any technique that enables computers to mimic human behavior.
    - *Example:* A robot with fixed instructions.
- **Machine Learning (ML):** A subset of AI. Systems that learn to perform a task from data without being explicitly programmed for every rule.
    - *Example:* A self-driving car.
- **Deep Learning (DL):** A subset of ML. It uses "deep" neural networks with many layers to learn from data. It's especially powerful for complex tasks like image recognition.

Moving Beyond Simple Models

- **Problem:** Can we use a simple model like **Logistic Regression (LR)** to classify complex medical images?
- **Answer:** Technically yes, but it would perform poorly (**underfitting**) because the relationship between pixels in an image is highly complex and non-linear. LR's model "capacity" is too low.
- **Solution: Deep Convolutional Neural Networks (CNNs)**, like Inception v3, which are designed to handle the complexity of image data through multiple layers of processing.

From LR to Deep Neural Networks (DNNs)

- To solve complex problems, we build **Deep Neural Networks** by:
  - Adding **Hidden Layers** and more **Nodes**.
  - Using **Non-Linear Activation Functions** (like Sigmoid or ReLU).
- **Universal Approximation Theorem:** A DNN with enough nodes/layers can approximate *any* continuous function, no matter how complex.
- **What do the internal nodes do?** They perform **feature extraction**. They combine input features (e.g., raw pixels) into new, more abstract features (e.g., edges, textures, shapes) that are useful for the final task (e.g., recognizing a dog's hoof). These learned features often don't have a simple human interpretation.

The Number of Parameters

- A key concept in deep learning is the **"number of parameters"** (weights and biases). This number grows very quickly as you add layers and nodes.
- **Significance:**
  - More parameters allow the network to learn more complex functions.
  - It also requires more data to train effectively and risks **overfitting** (memorizing the training data instead of learning general patterns).

Deep Neural Networks Summary

- A DNN has an input layer, multiple hidden layers, and an output layer.
- They are powerful, complex models with many parameters, suitable for solving very complex problems when you have a large amount of data.

Resources and Uncovered Topics

- Provides links for further learning on t-SNE, motifs, and UMAP.
- Mentions important topics for future lectures, like **Overfitting, Generalization, CNN (Convolutional Neural Networks)**